

# Optimal Lower Bound for Itemset Frequency Indicator Sketches

Eric Price  
UT Austin

October 13, 2014

## Abstract

Given a database, a common problem is to find the pairs or  $k$ -tuples of items that frequently co-occur. One specific problem is to create a small space “sketch” of the data that records which  $k$ -tuples appear in more than an  $\epsilon$  fraction of rows of the database.

We improve the lower bound of Liberty, Mitzenmacher, and Thaler [LMT14], showing that  $\Omega(\frac{1}{\epsilon}d\log(\epsilon d))$  bits are necessary even in the case of  $k = 2$ . This matches the sampling upper bound for all  $\epsilon \geq 1/d^{.99}$ , and (in the case of  $k = 2$ ) another trivial upper bound for  $\epsilon = 1/d$ .

## 1 Introduction

[Check out [LMT14] for a more complete introduction.]

We are concerned with sketches for itemset frequencies in databases. The “itemset frequency” is the fraction of rows in a database where a set of items co-occur:

**Definition 1.1** (Itemset Frequency). *For a database  $\mathcal{D} \in (\{0, 1\}^d)^n$  and an itemset  $T \subseteq [d]$ , the frequency of  $T$  in  $\mathcal{D}$  is*

$$f_T(\mathcal{D}) = \frac{1}{n} |\{i : \forall j \in T, (\mathcal{D}_i)_j = 1\}|$$

An itemset frequency indicator sketch is a smaller space representation of  $\mathcal{D}$  that lets us identify the itemsets with large frequency:

**Definition 1.2** (Itemset-Frequency-Indicator sketches). *An Itemset-Frequency-Indicator sketching scheme is a pair of algorithms: one receives  $k, \epsilon$  and a database  $\mathcal{D} \in (\{0, 1\}^d)^n$  and outputs a sketch  $S \in \{0, 1\}^m$ , and another takes  $S, \epsilon$ , and a set  $T \subset [d]$  with  $|T| = k$ , and returns an estimate of whether  $f_T(\mathcal{D}) > \epsilon$ . In particular, it must output YES if*

$$f_T(\mathcal{D}) \geq \epsilon$$

*and NO if*

$$f_T(\mathcal{D}) \geq \epsilon/2.$$

*For this problem, we require that the first algorithm “succeed” with 3/4 probability, and if it does then the second algorithm should always output the correct answer for every query  $T$ .*

The question is: how large must  $m$  to solve this problem? If we allowed the queries to fail with a small constant probability, then per [LMT14] the space complexity is  $\Theta(d/\epsilon)$ . The goal of this paper is to get an extra  $\log d$  factor from needing to union bound over  $d^k$  queries.

There are two trivial upper bounds, for constant  $k$ :

- Sampling takes  $O(\frac{1}{\epsilon}d \log d)$  bits of space.
- Storing all the answers takes  $O(d^k)$  bits of space.

We show that  $\Omega(\frac{1}{\epsilon}d \log(\epsilon d))$  bits are necessary even in the case of  $k = 2$ . This means that sampling is optimal for all  $\epsilon \geq 1/d^{1-\alpha}$  for any constant  $\alpha > 0$ , while storing all answers is optimal for  $\epsilon \leq 1/d$  and  $k = 2$ .

**Theorem 3.2.** *Any sketch for the Itemset-Frequency-Indicator problem must take  $\Omega(\frac{1}{\epsilon}d \log(\epsilon d))$  space for all  $1/d \leq \epsilon \leq 1/8$ , even in the case of  $k = 2$ .*

For  $k = 2$ , in the relatively minor intermediate regime of  $\epsilon = 1/d^{1-o(1)}$ , it seems likely that neither trivial upper bound is quite tight. For  $k > 2$ , one can probably extend the result to show that sampling is optimal for  $\epsilon > 1/d^{k-1-\alpha}$ ; we leave these questions to future work.

A more interesting open question is for itemset frequency estimation. If we want to estimate  $f_T(\mathcal{D})$  to  $\pm\epsilon$ , then sampling requires  $O(\frac{1}{\epsilon^2}d \log d)$  space but we don't know any better lower bound than the above  $\Omega(\frac{1}{\epsilon}d \log d)$  bound. ([LMT14] first showed this for  $1/d^{1-\alpha} \ll \epsilon \ll 1/\log d$ , and our Theorem 3.2 removes the upper limit on  $\epsilon$ ).

To the best of our knowledge, [LMT14] contains the only previous space lower bound for this type of problem. A number of other aspects of the problem have been studied, however; see [LMT14] for an overview of related work. Our theorem is a strict improvement over their Theorem 18, which gets  $\Omega(\frac{1}{\epsilon^{1-1/k}}d \log d)$  for a restricted range of  $\epsilon$ .

## 2 Notation

We use  $[n]$  to denote  $\{1, 2, \dots, n\}$ . For two vectors  $v \in \mathbb{R}^d$  and  $w \in \mathbb{R}^{d'}$ , we use  $v \parallel w$  to denote the  $d + d'$  dimensional vector that is the concatenation of  $v$  and  $w$ .

## 3 Proof

For simplicity of exposition, we begin with the  $\epsilon = \Theta(1)$  case, which was not previously known ([LMT14] required  $\epsilon \ll 1$ ). The general  $\epsilon$  case follows a very similar outline.

**Lemma 3.1.** *Any sketch for the Itemset-Frequency-Indicator problem with  $\epsilon = 1/8$  must take  $\Omega(d \log d)$  space.*

*Proof.* Let  $m = d/2$ . We will encode an arbitrary permutation  $\Pi$  of  $[m]$  into the results of Itemset-Frequency-Indicator. This forces Itemset-Frequency-Indicator to store at least  $\log(m!) = \Theta(m \log m) = \Theta(d \log d)$  bits.

For each  $i$ , define  $e_i \in \{0, 1\}^m$  to be the elementary unit vector with a 1 in position  $i$ . Given a subset  $S$  of  $[m]$ , we associate a vector

$$v_S := \left( \sum_{i \in S} e_i \right) \parallel \left( \sum_{i \in \overline{S}} e_{\Pi(i)} \right)$$

where  $\parallel$  denotes concatenation and  $\overline{S} = [m] \setminus S$ .

Our database simply consists of  $n = \Theta(\log d)$  vectors  $v_S$  for independent, randomly chosen  $S$ . In particular, each  $S$  contains each element of  $[m]$  with probability  $1/2$ .

Now, for each row  $v_S$  and any  $i, j \in [m]$  consider the distribution on the co-occurrence of the itemset  $\{i, m + j\}$ . If  $j = \Pi(i)$ , this conjunction never appears. If  $j \neq \Pi(i)$ , on the other hand, then the conjunction appears with  $1/4$  probability.

After looking at  $n = \Theta(\log d)$  such vectors, with high probability all itemsets  $\{i, m + j\}$  with  $j \neq \Pi(i)$  will have more than  $n/8$  appearances. Then  $f_{\{i, m+j\}}(\mathcal{D})$  will be 0 if  $j = \Pi(i)$  and at least  $1/8$  if  $j \neq \Pi(i)$ . Therefore an  $\epsilon = 1/8$  Itemset-Frequency-Indicator algorithm will return NO for  $\{i, m + j\}$  precisely when  $j = \Pi(i)$ , so we can recover  $\Pi$  from the sketch. Hence the sketch must have  $\Omega(d \log d)$  bits.  $\square$

We now extend this approach to general  $\epsilon$  with  $1/d \leq \epsilon \leq 1$ .

**Theorem 3.2.** *Any sketch for the Itemset-Frequency-Indicator problem must take  $\Omega(\frac{1}{\epsilon} d \log(\epsilon d))$  space for all  $1/d \leq \epsilon \leq 1/8$ , even in the case of  $k = 2$ .*

*Proof of Theorem 3.2.* Let  $m = \epsilon d/2$ , which we can assume is an integer by rescaling constants. We will encode  $1/\epsilon^2$  permutations  $\Pi_{k,l}$  of  $[m]$ , for  $k, l \in [1/\epsilon]$ . This requires  $(1/\epsilon^2) \log((\epsilon d/2)!) = \Theta(\frac{1}{\epsilon} d \log(\epsilon d))$  bits, giving the result.

Let  $e_i \in \{0, 1\}^m$  denote the elementary unit vector with a 1 in position  $i$ . For any  $S \subset [m]$  and  $k \in [1/\epsilon]$ , we first define  $u^{k,S} \in \{0, 1\}^{d/2}$  by

$$u_i^{k,S} = 1 \text{ if and only if } i = (k-1)m + j \text{ for some } j \in S$$

to represent the set  $S$  in “block”  $k$ . We then define the associated vector  $v_{k,S} \in \{0, 1\}^d$  by

$$v_{k,S} := u^{k,S} \parallel \left( \sum_{i \in \overline{S}} e_{\Pi_{k,1}(i)} \right) \parallel \left( \sum_{i \in \overline{S}} e_{\Pi_{k,2}(i)} \right) \parallel \cdots \parallel \left( \sum_{i \in \overline{S}} e_{\Pi_{k,1/\epsilon}(i)} \right).$$

We then choose  $n = \Theta(\frac{1}{\epsilon} \log d)$  vectors for the database by, for each  $k \in [1/\epsilon]$ , choosing  $\Theta(\log d)$   $v_{k,S}$  for uniformly random  $S \subseteq [m]$ .

Given the database, to figure out  $\Pi_{k,l}(i)$  we query the itemset  $T_{k,l}(i, j) = \{(k-1)m + i, d/2 + (l-1)m + j\}$  for all  $j \in [m]$ . We have that  $T_{k,l}(i, j)$  appears in  $v_{k',S}$  exactly when  $k' = k$  with  $i \in S$  and  $\Pi_{k,l}^{-1}(j) \notin S$ . Thus it never appears if  $j = \Pi_{k,l}(i)$ , but otherwise it appears in each sampled  $v_{k,S}$  with probability  $1/4$ . Thus with high probability, it will appear in at least  $\epsilon n/8$  of the rows. By a union bound, with high probability  $f_{T_{k,l}(i,j)}(\mathcal{D}) \geq \epsilon/8$  for all  $i, j, k, l$  with  $j \neq \Pi_{k,l}(i)$ , while it is zero when  $j = \Pi_{k,l}(i)$ . Hence an  $\epsilon/8$ -approximate solution to Itemset-Frequency-Indicator would let us recover all the  $\Pi_{k,l}$  with high probability, retrieving  $\Theta(\frac{d}{\epsilon} \log(\epsilon d))$  bits of information. Therefore the sketch must store this many bits.  $\square$

## References

- [LMT14] Edo Liberty, Michael Mitzenmacher, and Justin Thaler. Space lower bounds for itemset frequency sketches. *arXiv preprint arXiv:1407.3740*, 2014.